Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 06

Pierre-Luc Germain

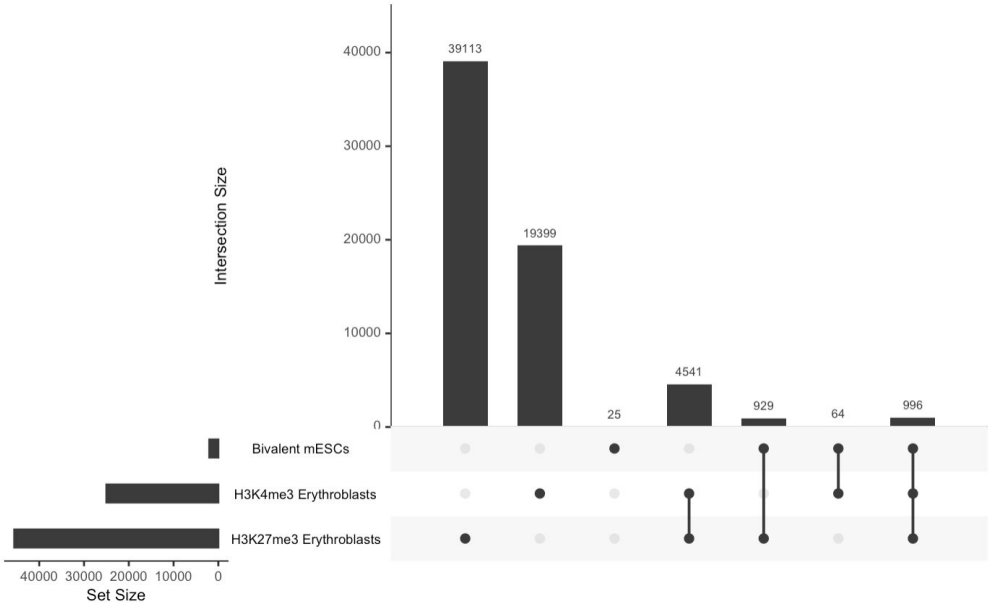
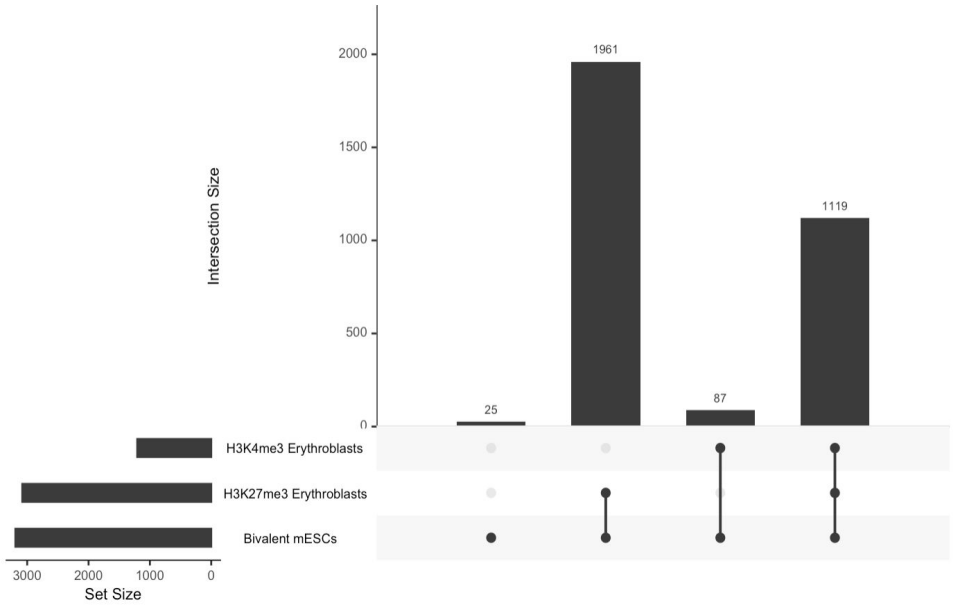
ETH Zürich

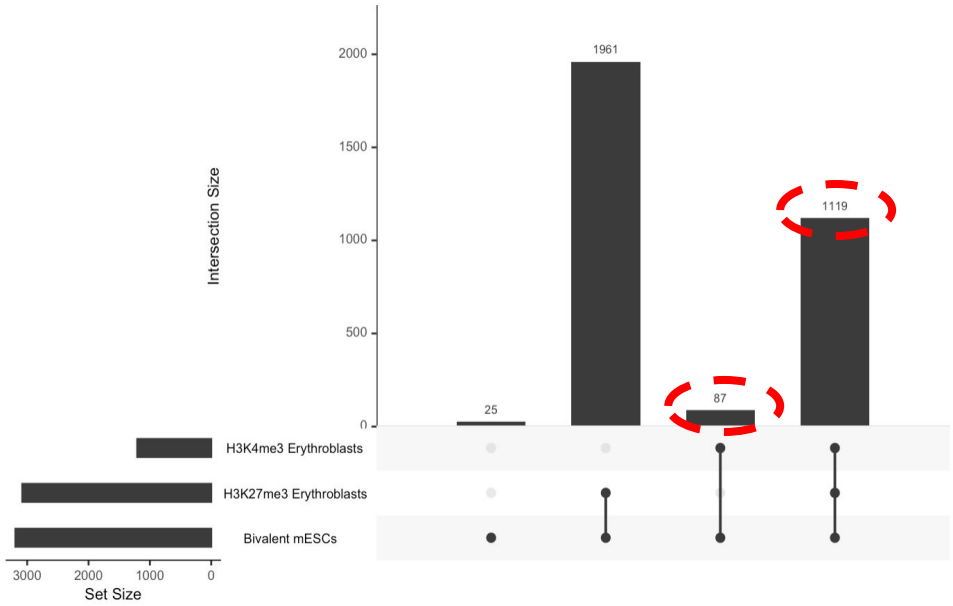
Plan

- New packages to install (see slack)
- Debriefing on last week’s assignment
- Overview of transcription factors and their binding specificity
- DNA motifs and related analysis

Debriefing on the assignments

Using references for upset plot

```
```{r, without reference}
# without reference
peakList <- list(biValMe_2, H3K4me3_eb, H3K27me3_eb)
names(peakList) <- c("Bivalent mESCs", "H3K4me3 Erythroblasts", "H3K27me3 Erythroblasts")
regionUpset(peakList)
```
```



Debriefing on the assignments

Using references for upset plot

```
```{r, with reference}
# with reference
regionUpset(peakList, reference=peakList[[1]])
```
```



Debriefing on the assignments

Using references for upset plot

```
```{r, with reference}
# with reference
regionUpset(peakList, reference=peakList[[1]])
```
```



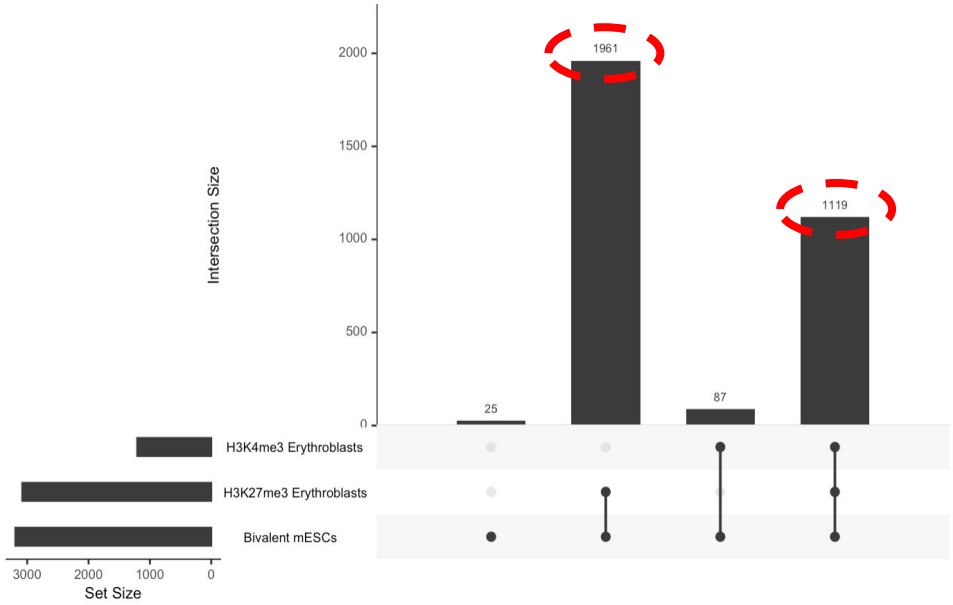

```
> sum(overlapsAny(biValMe_2, H3K4me3_eb))
[1] 1206
```

=87+1119

Debriefing on the assignments

Using references for upset plot

```{r, with reference}
# with reference
regionUpset(peakList, reference=peakList[[1]])
```

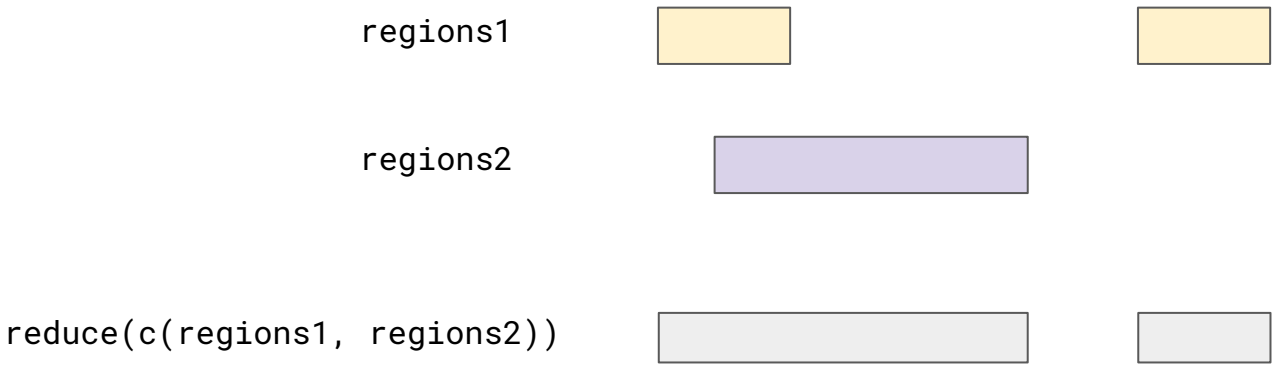



```
> sum(overlapsAny(biValMe_2, H3K27me3_eb))
[1] 3080
```

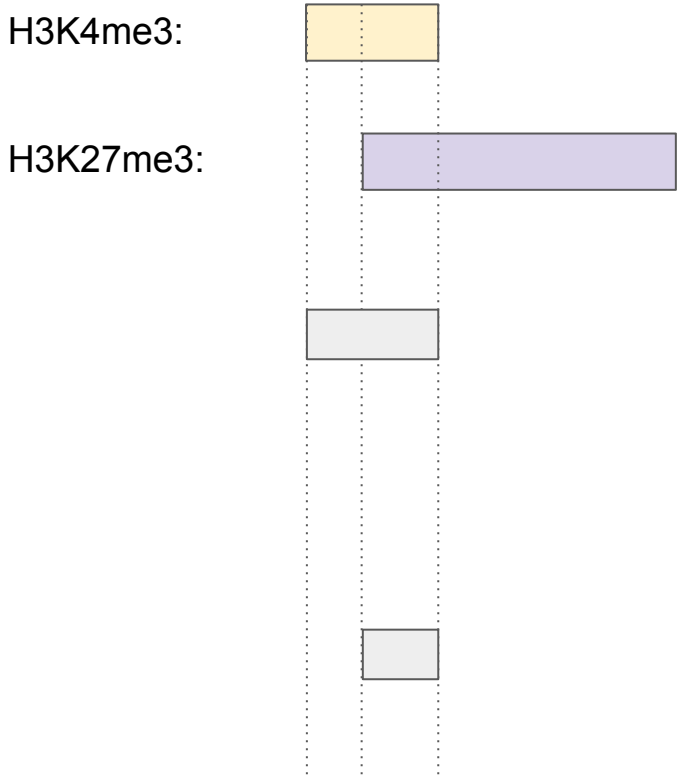
=1916+1119

Debriefing on the assignments

When no reference is specified, one is created automatically by merging and *reducing* the regions (unless otherwise specified in the arguments):



Intersection & overlap:
The example of bivalent domains



- **method one (overlapsAny):**
  find the H3K4me3 peaks that overlap a H3K27me3 domain

- **method two (intersect):**
  find the regions that are covered by both H3K4me3 and H3K27me3

Debriefing on the assignments

Annotations:

ENCFF247GVM

Histone ChIP-seq in ES-Bruce4

Mus musculus strain Bruce4 ES-Bruce4

Target: H3K4me3
Lab: Bing Ren, UCSD
Project: ENCODE
Reference Epigenome: ENCSR343RKY
candidate Cis-Regulatory Elements (cCREs): SCREEN

ENCFF326VMV

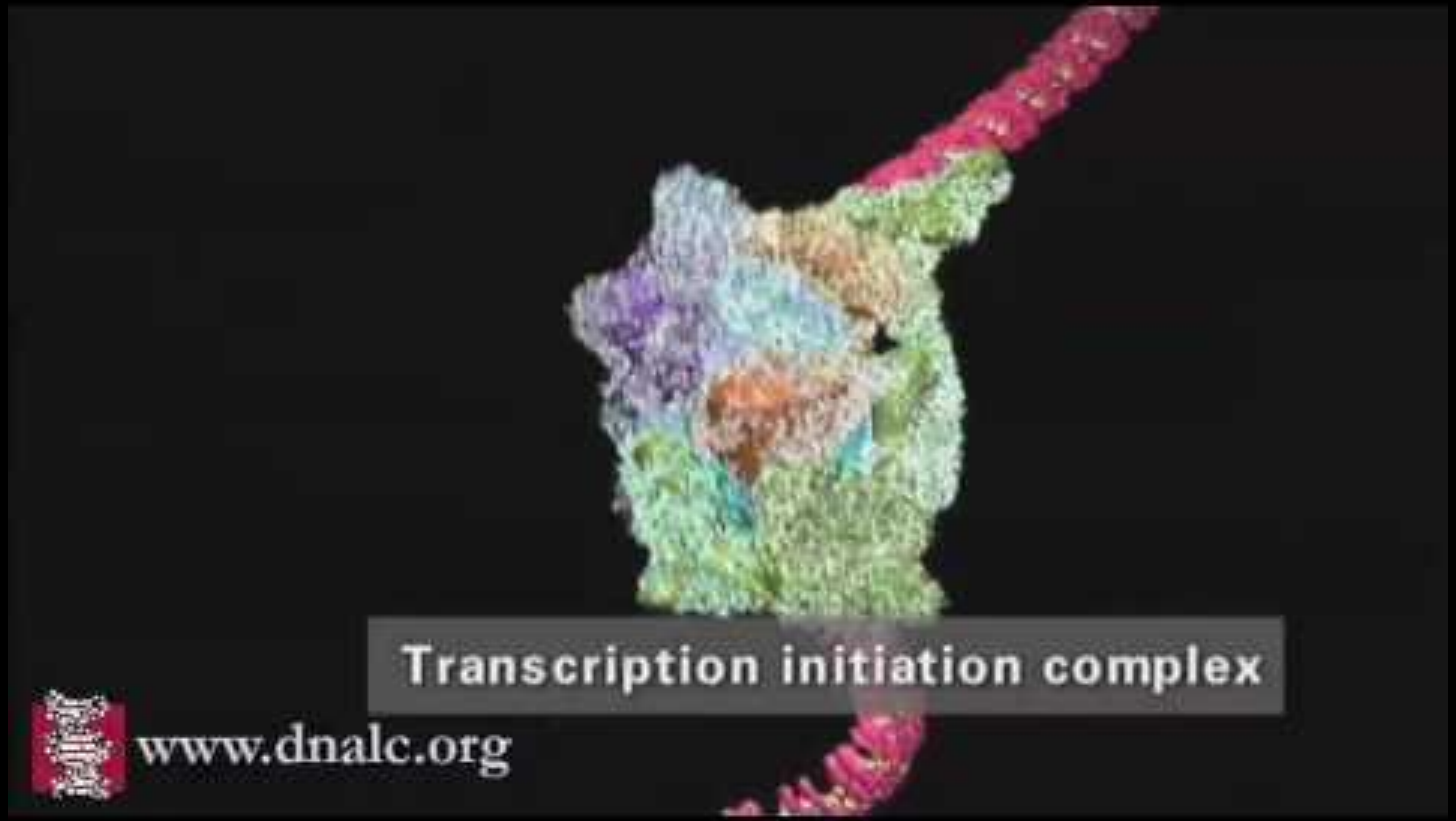
Histone ChIP-seq in smooth muscle cell

Homo sapiens smooth muscle cell originated from H9

Target: H3K4me3
Lab: Bradley Bernstein, Broad
Project: ENCODE
Reference Epigenome: ENCSR116JEF
candidate Cis-Regulatory Elements (cCREs): SCREEN

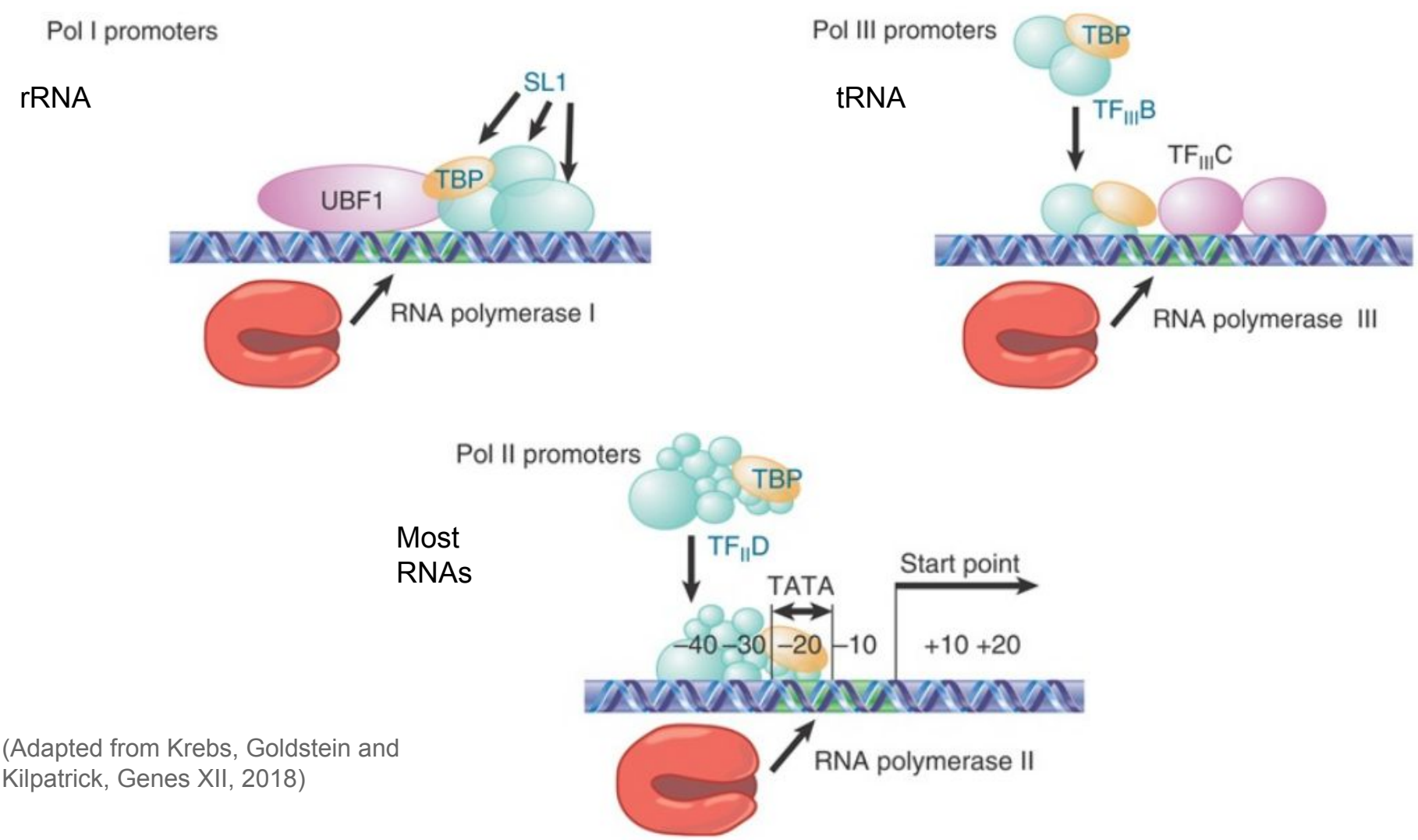
```
regionUpset(peaks, nsets=length(peaks))
```

```
## Warning in .merge_two_Seqinfo_objects(x, y): Each of the 2 combined objects has sequence levels not in the oth
er:
##   - in 'x': chr20, chr21, chr22
##   - in 'y': chr4_GL456216_random, chrUn_GL456368, chrUn_GL456370, chrUn_GL456378, chrUn_JH584304, chrX_GL45623
3_random
##   Make sure to always combine/compare objects based on the same reference
##   genome (use suppressWarnings() to suppress this warning).
```



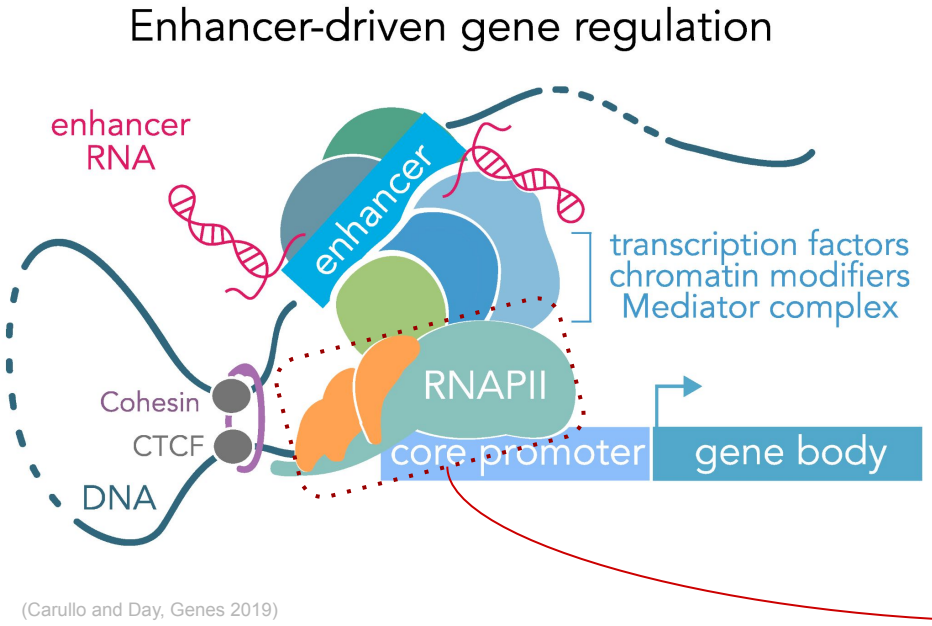
https://youtu.be/SMtWvDbfHLo

(See also https://youtu.be/WW9lIYM_FC0)

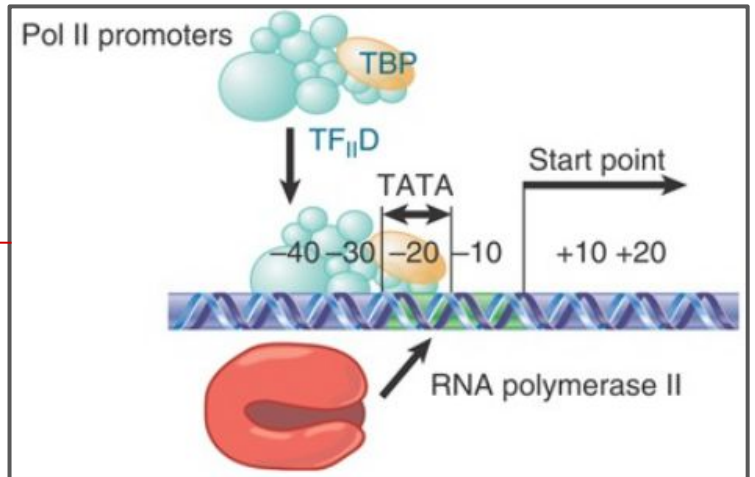


(Adapted from Krebs, Goldstein and Kilpatrick, Genes XII, 2018)

Additional regulatory elements



(Carullo and Day, Genes 2019)



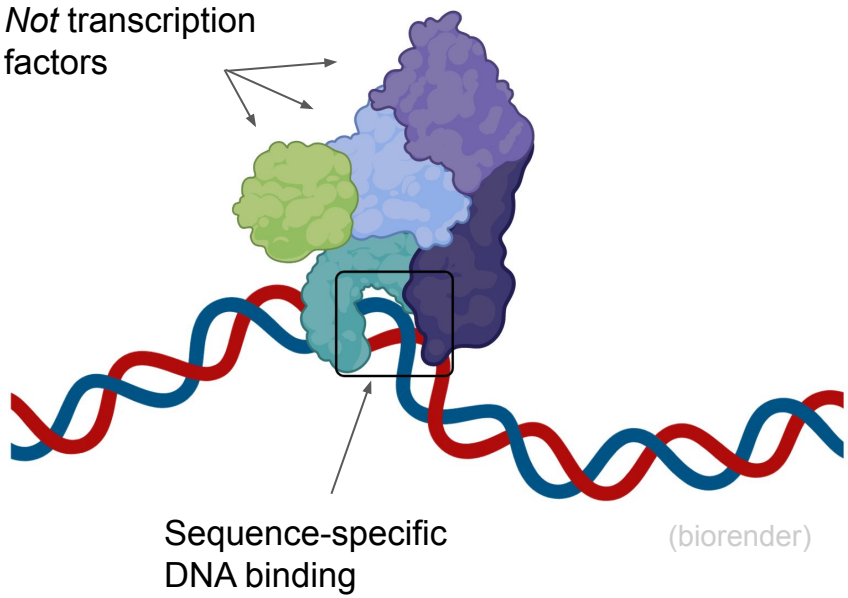
What is a transcription factor?

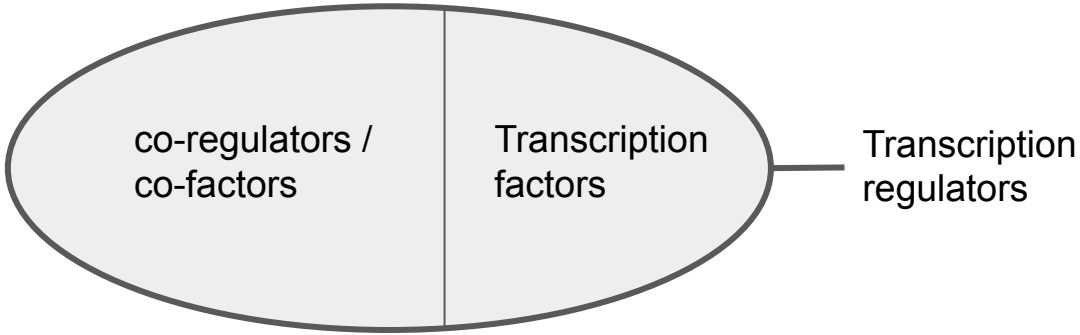
Proteins capable of both:

1) Binding DNA in a sequence-specific manner

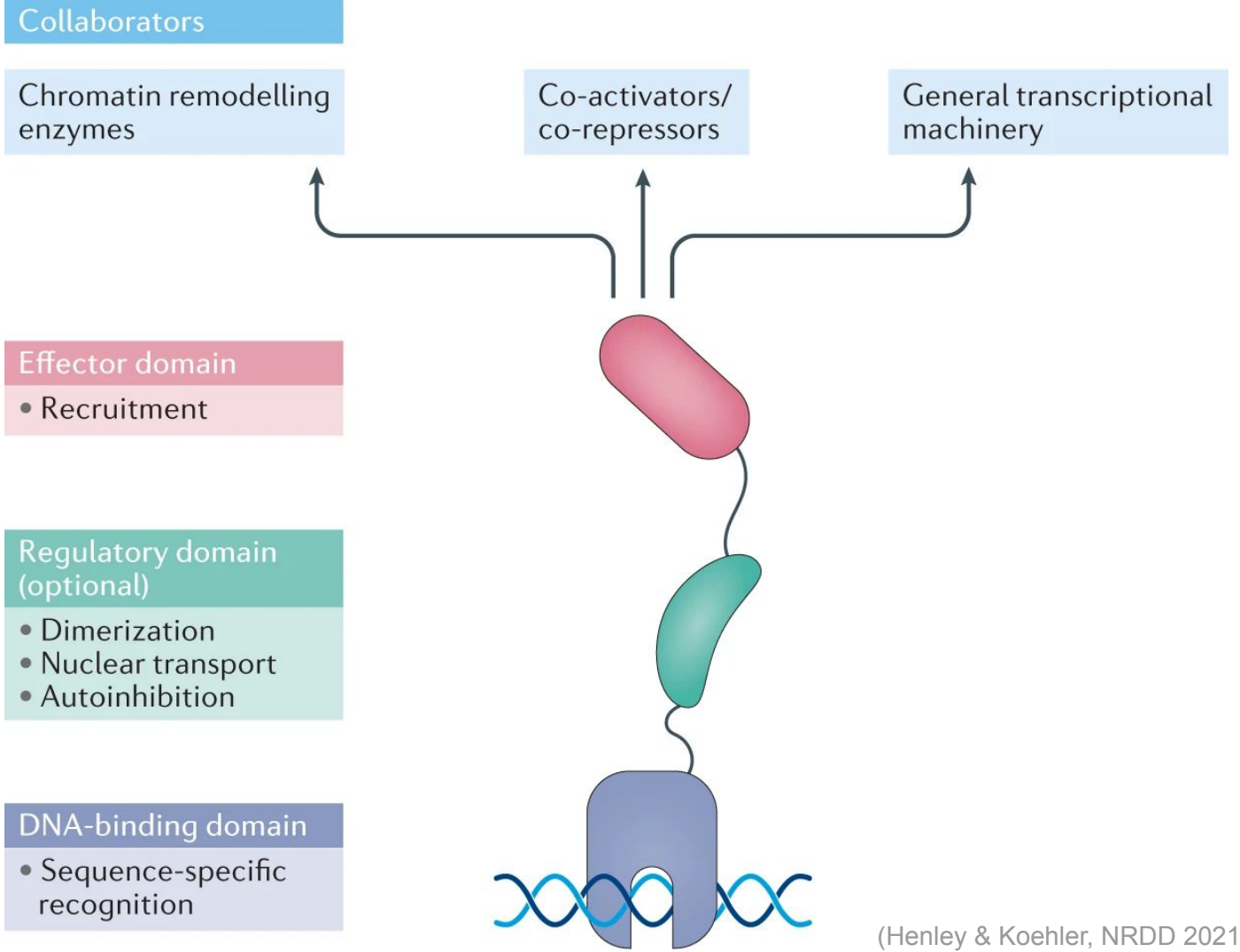
2) Regulating transcription

(Lambert et al., Cell 2018)

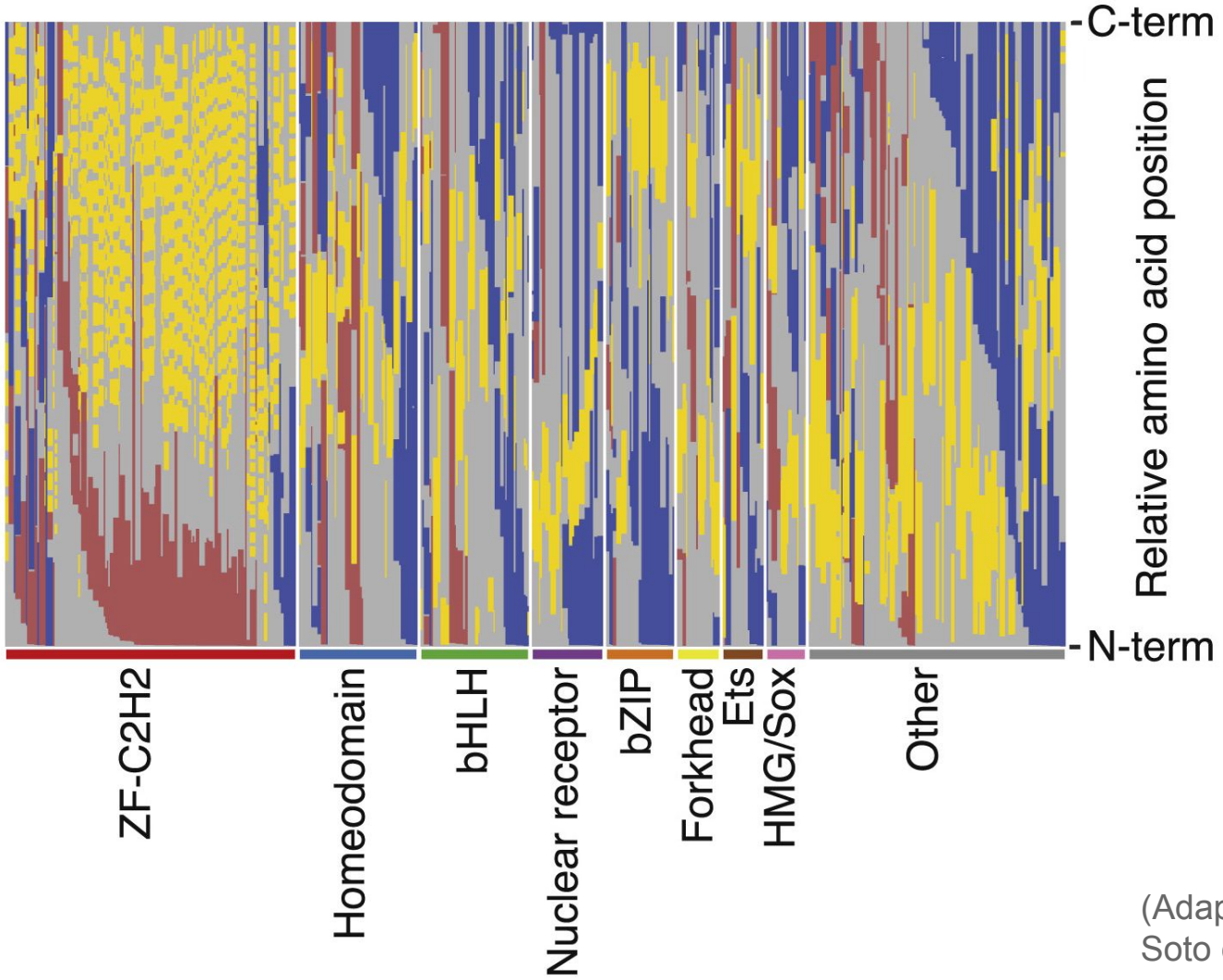




Anatomy of a transcription factor (TF)



(Henley & Koehler, NRDD 2021)



While most TF have either an activating (AD) or repressive (RD) domain, some have both

Domain
AD
RD
DBD
Other

(Adapted from
Soto et al., Molecular Cell 2021)

Review (Cell 2018)

The Human Transcription Factors

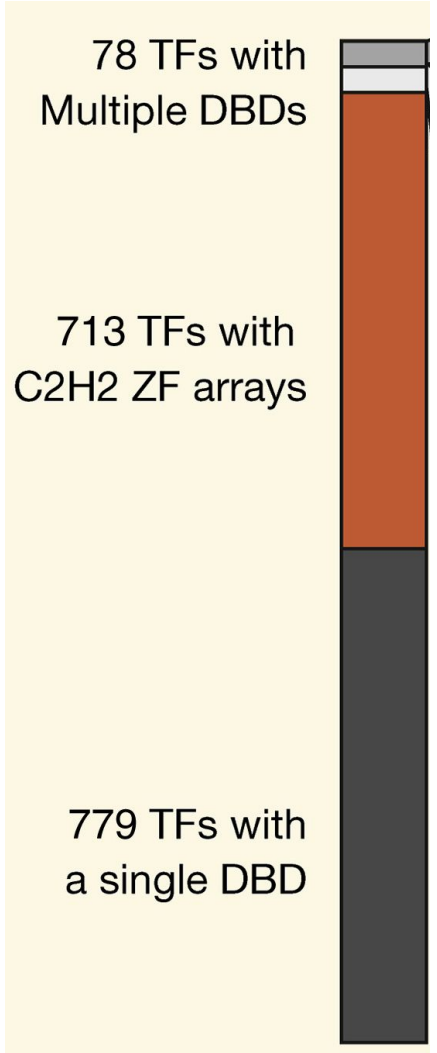
Samuel A. Lambert [1, 9], Arttu Jolma [2, 9], Laura F. Campitelli [1, 9], Pratyush K. Das [3], Yimeng Yin [4], Mihai Albu [2], Xiaoting Chen [5], Jussi Taipale [3, 4, 6] ✉, Timothy R. Hughes [1, 2] ✉, Matthew T. Weirauch [5, 7, 8] ✉

Proteins capable of both:

1) Binding DNA in a sequence-specific manner

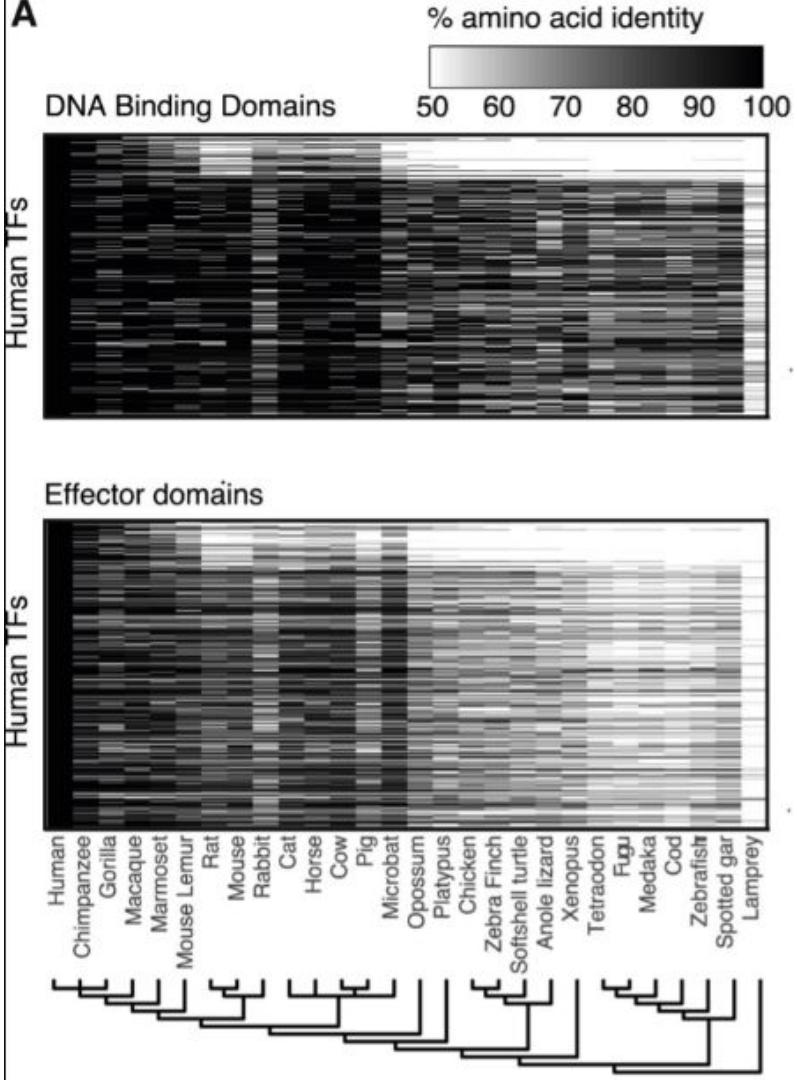
2) Regulating transcription

According to their census, humans have 1570 transcription factors

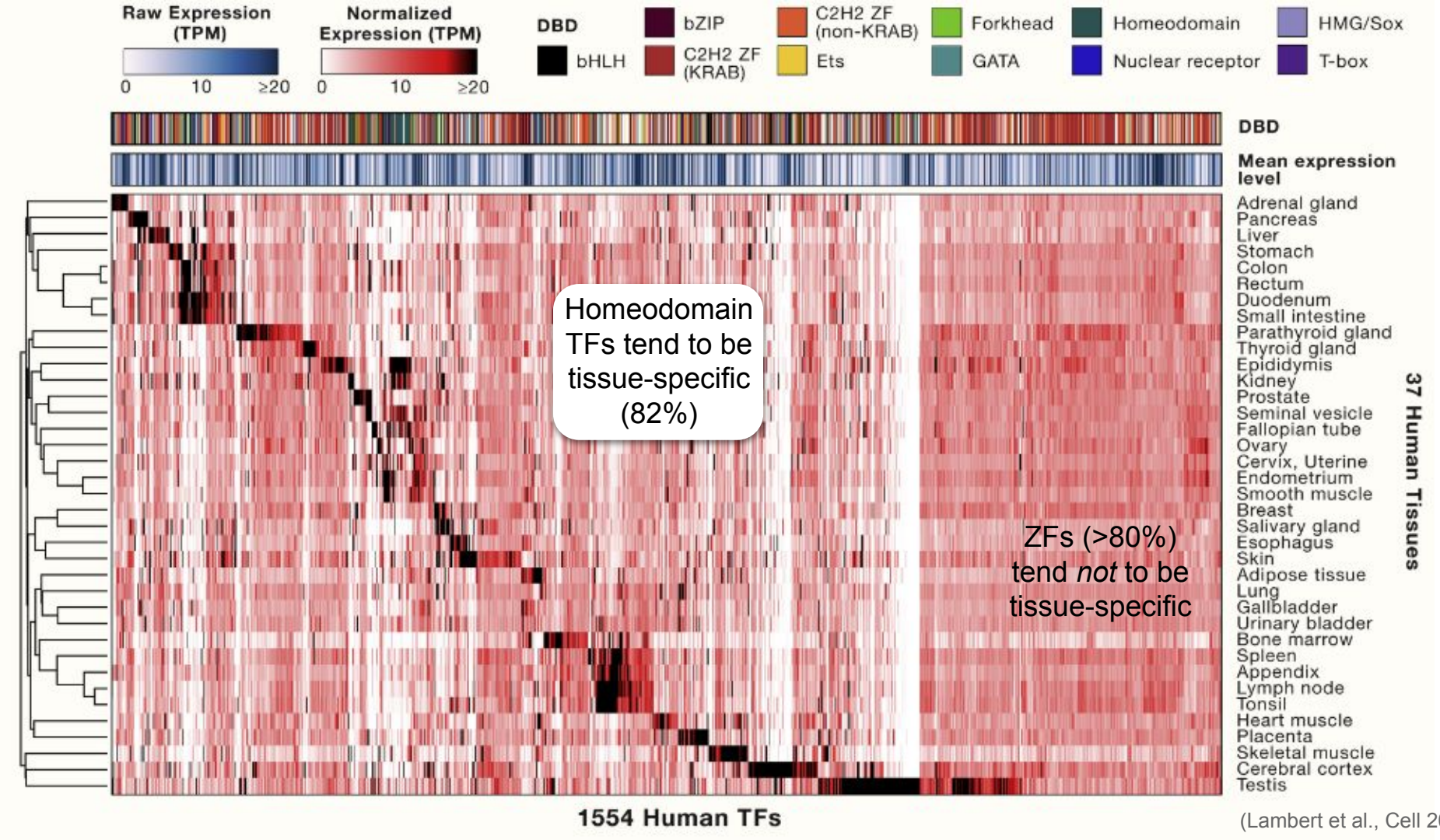


Transcription factors are highly conserved

DNA binding domains show much higher conservation than effector domains



(Soto et al., Molecular Cell 2021)

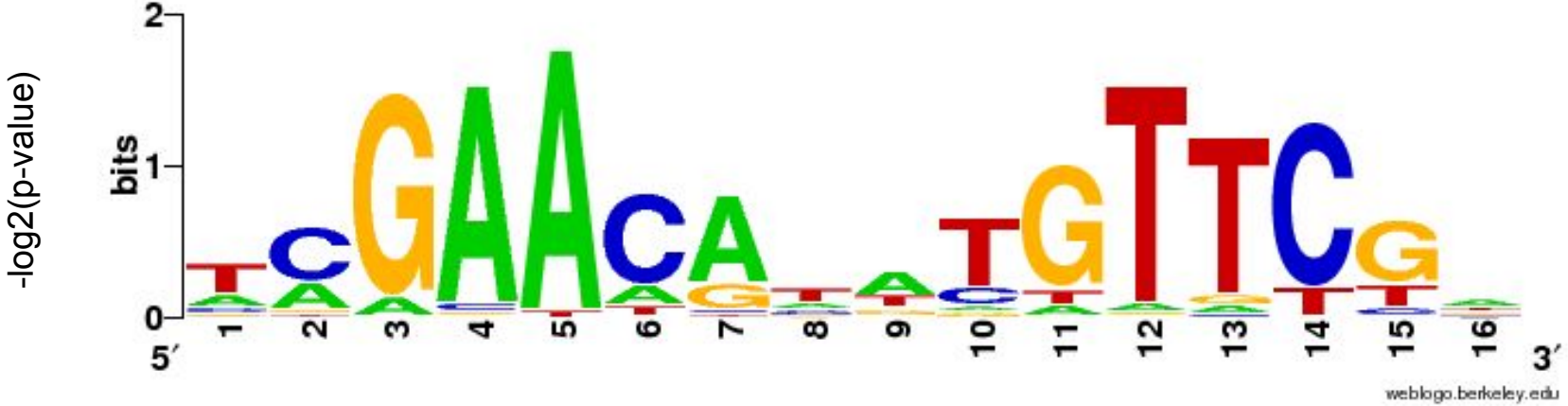


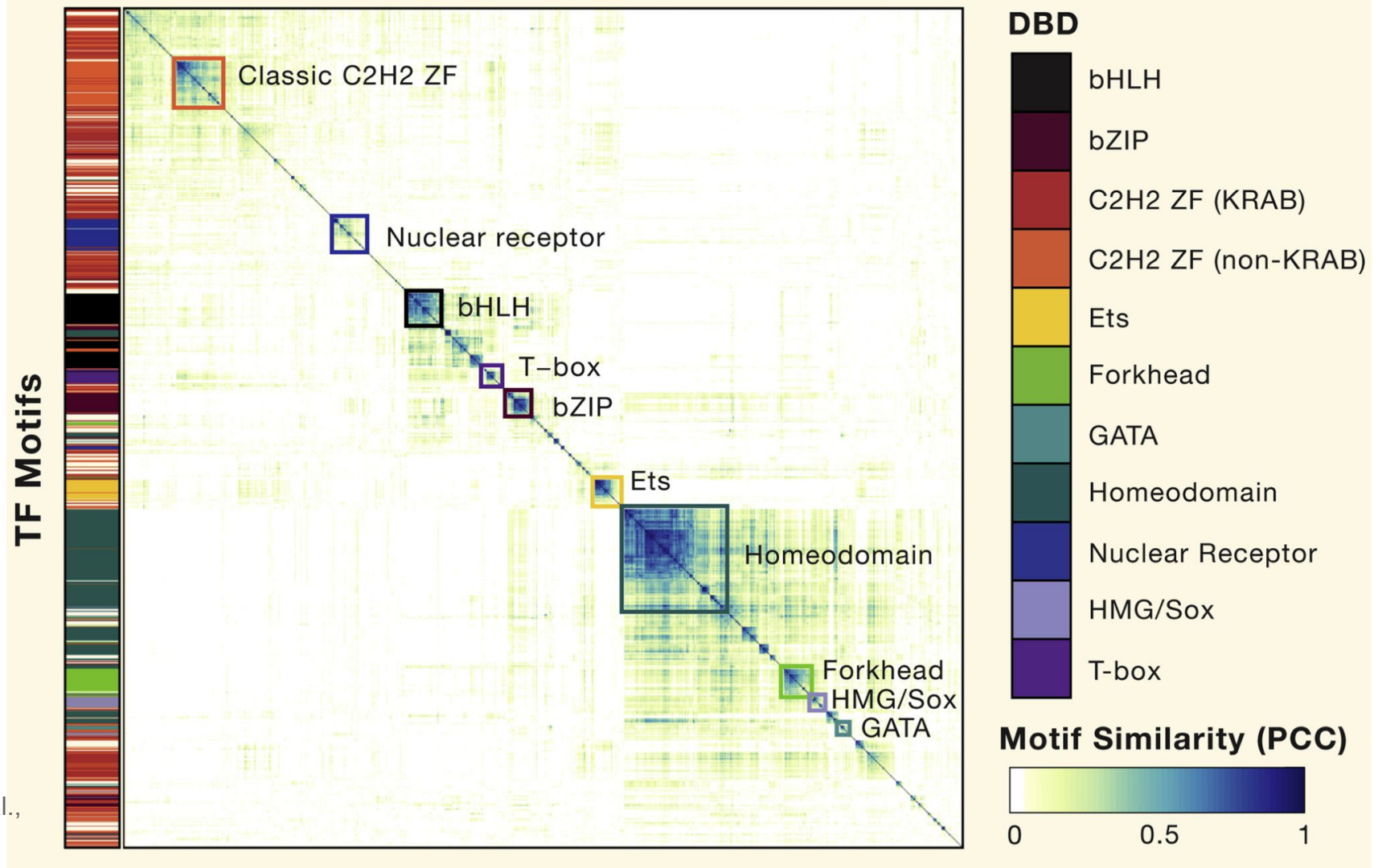
(Lambert et al., Cell 2018)

Sequence-specificity

E.g. The LexA bacterial TF recognizes the consensus sequence

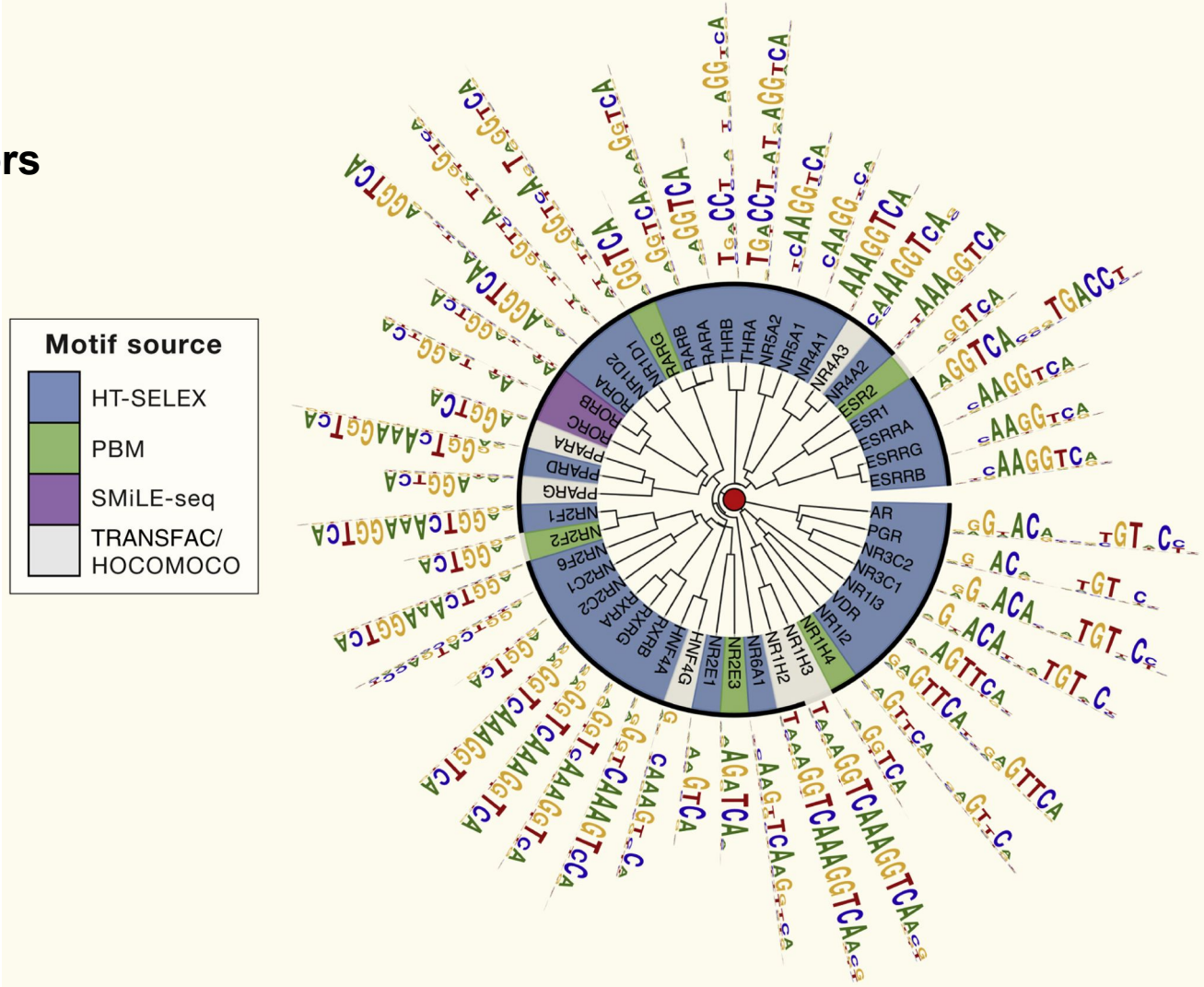
5'-GAACAnnTGTTC-3'





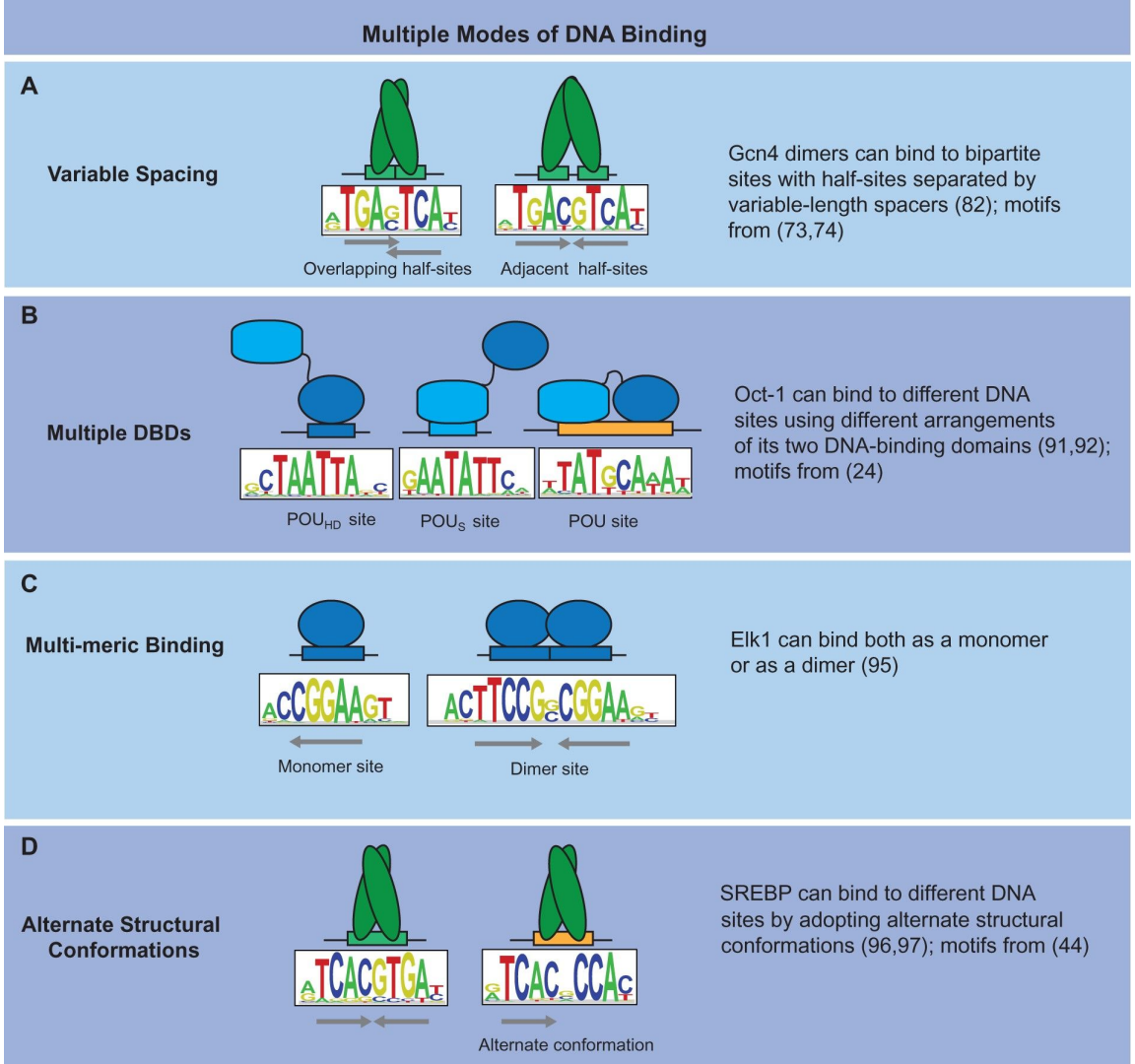
(Lambert et al., Cell 2018)

An example of TF motif degeneracy: Nuclear hormone receptors



(Lambert et al., Cell 2018)

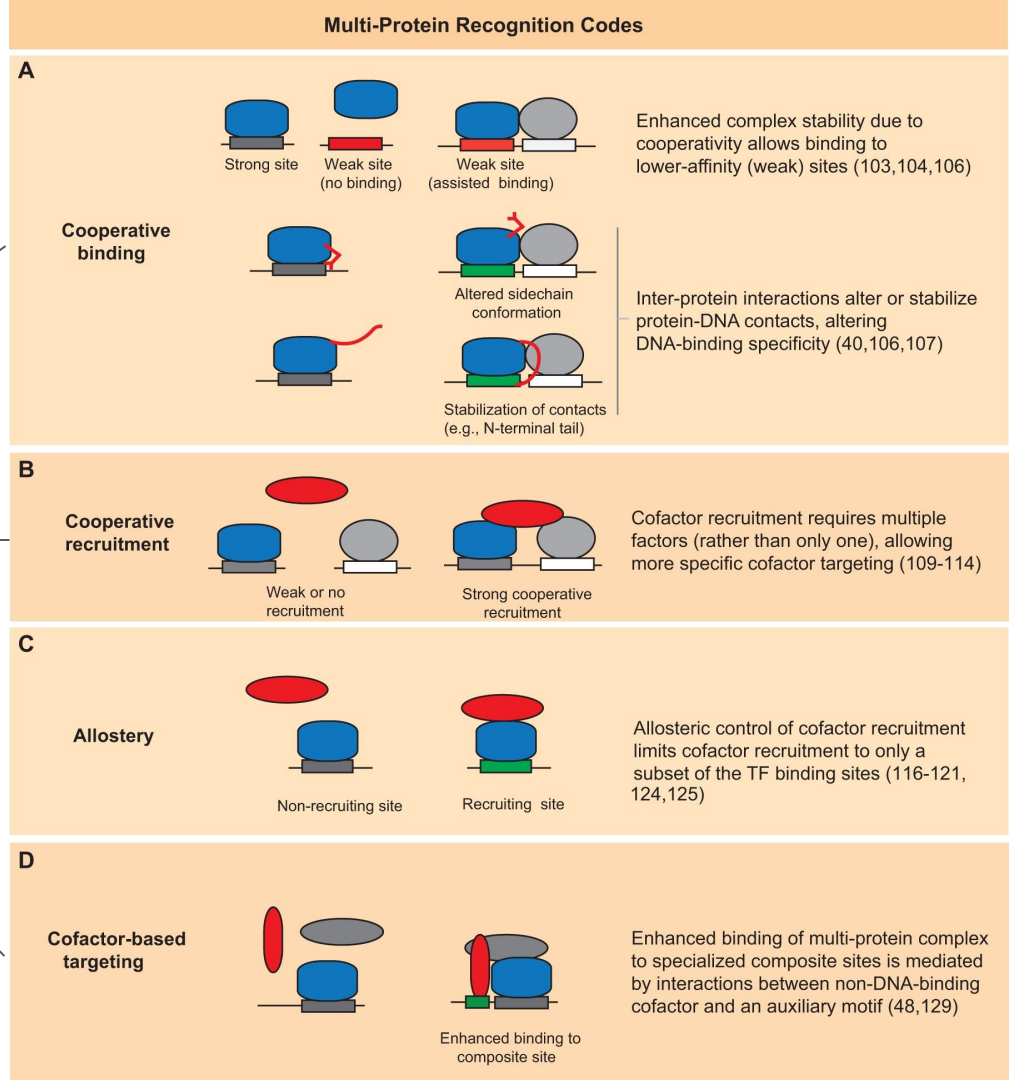
Variations in DNA binding specificity



(Siggers and Gordân, NAR 2014)

Cooperative binding

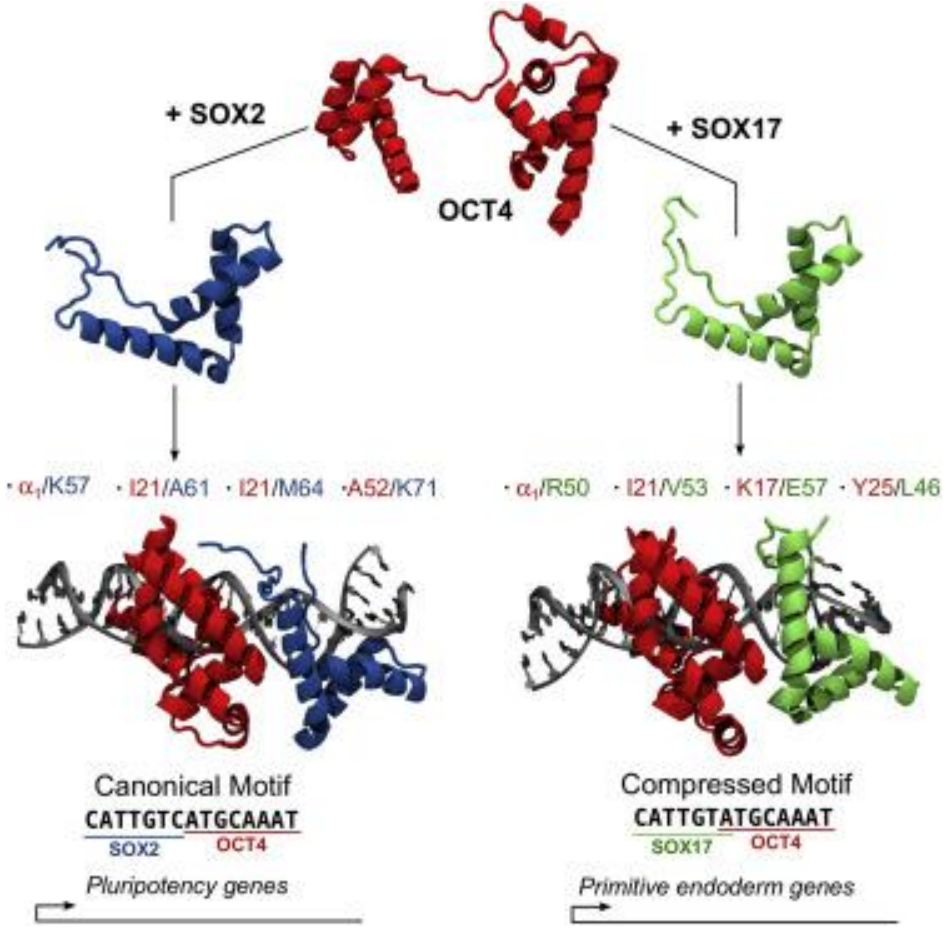
Highly combinatorial binding of TFs



(Siggers and Gordân, NAR 2014)

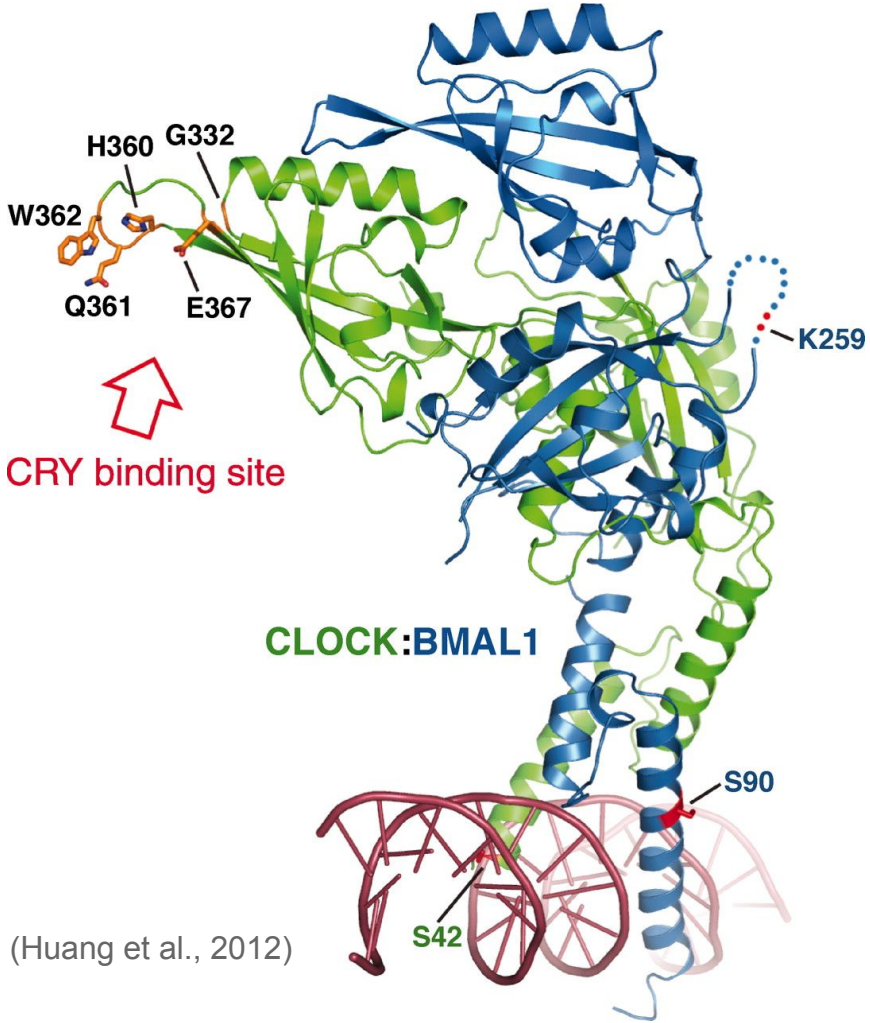
Two examples of Cooperative binding

OCT4 (POU5f1) binding upon differentiation

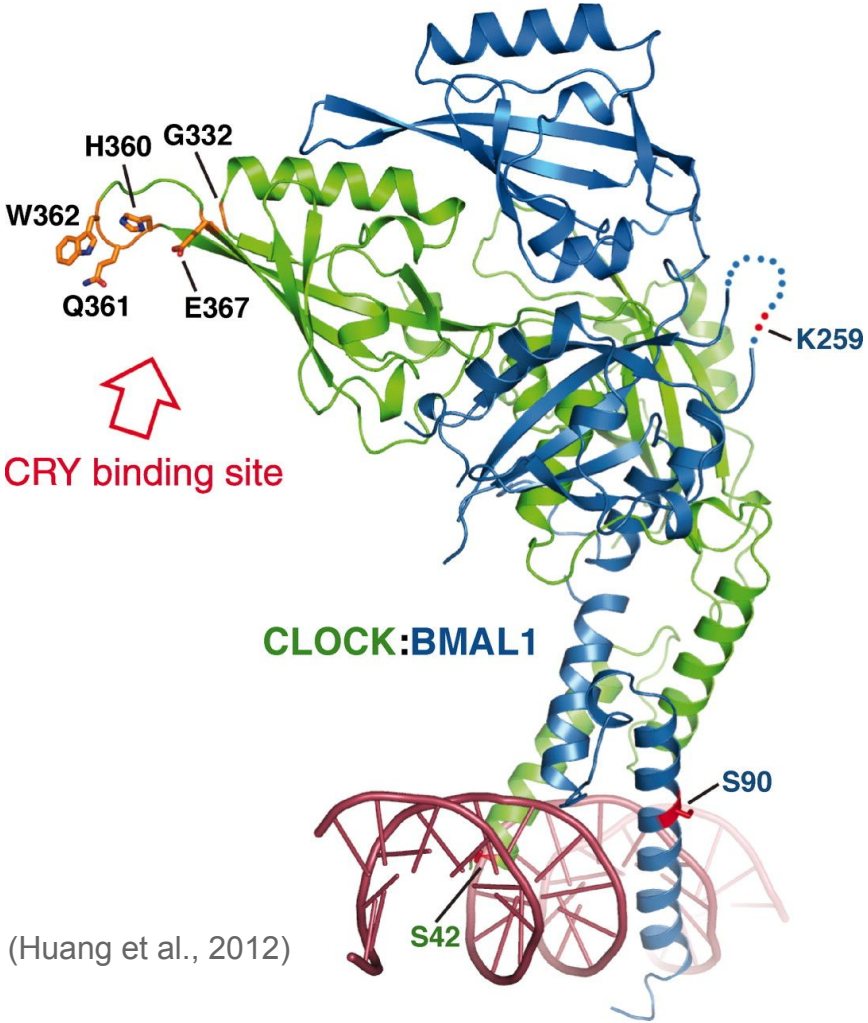


(Merino et al., *Structure* 2014)

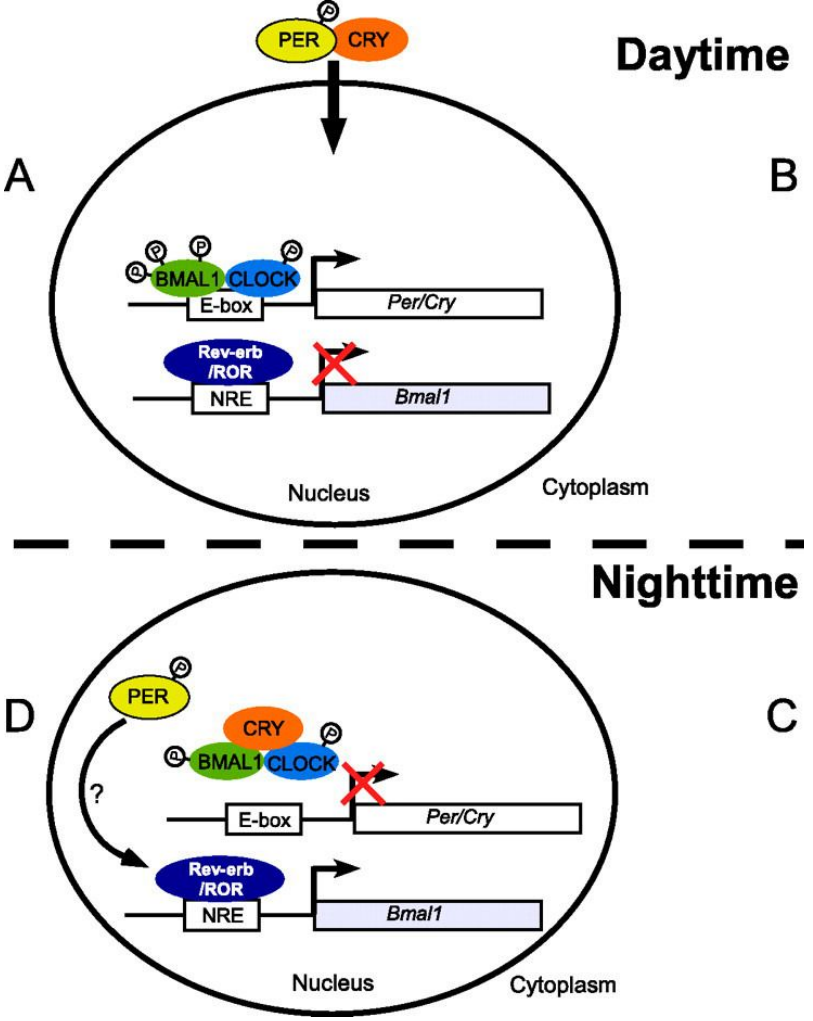
Clock-Bmal-Cry during circadian rythm



(Huang et al., 2012)



(Huang et al., 2012)



(Adapted from Ye et al., JBC 2011)

Motif analysis

- **Motif discovery** aims at finding **new** motifs that are enriched in a set of sequences (e.g. peaks) versus a background
  - Example method: meme (Meme suite)
  - Bioconductor method: `rGADEM` package (see also the `memes` R package)

- **Motif enrichment** analysis aims at finding **known** motifs that are enriched in a set of sequences (e.g. peaks) versus a background
  - Example method: AME (Meme suite)
  - Bioconductor method: `PWMEnrich` package

- **Motif scanning** aims at finding the **occurrences of known** motifs in a set of sequences (methodologically fairly simple – which method doesn't matter much)
  - Example method: fimo (Meme suite)
  - Bioconductor method: `motifmatchr` (see also `TFBSTools` package)

Genetic variation at TF binding sites

- Genetic variation at TF binding sites can affect the binding of the protein, and hence impact development and health
- Nevertheless, while most coding sequences show evidence of **evolutionary constraint** (e.g. purifying selection), only a small fraction of TF binding sites (11.6% of footprints) show evidence of constraint – the vast majority appears to be evolving neutrally

(Vierstra et al., Nature 2020)

- This suggests a degree of (at least partial) redundancy between regulatory elements

Assignment

- Choose a transcription factor, e.g. CREB1, REST, GATA5, EGR1, GCR
  (or any of your choice that has a motif and available ChIPseq data)
- Download the (e.g. Mouse) peaks for that factor (whatever cell type)
- Identify the instances of the factor's motif
- Answer the following questions:
  - Of all the peaks, what proportion contains a motif for the factor?
    - Expected form of an answer: of the XX peaks, XX (XX%) contain a motif
  - Of all instances of that motif in the genome (or in one chromosome), what proportion is bound by the factor (i.e. has a peak)?
    - Expected form of an answer: of the XX motif instances, XX (XX%) overlap a peak

Don't forget to *render* your markdown and push it as `assignment.html` !